

# **Practical approaches to standardizing vocabularies: the Cultural Heritage experience.**

Phil Carlisle

English Heritage National Monuments Record

and

European Heritage Network



# Introduction

- Introduction
- Controlled Vocabularies
- Case Study

# Introduction

- About me
- Data Standards Supervisor for the National Monuments Record (RCHME/EH)
- Over 10 years experience of constructing controlled vocabularies for the Heritage sector in the UK
- Editor of the UK Archival Thesaurus
- English language representative on the HEREIN Thesaurus 'Task Force'



# Controlled Vocabularies

- Wordlists
  - Castle
  - Ditch
  - House
  - Terraced House
- Complex Wordlists
  - Dwellings
    - House
    - Bungalow
  - Defensive Buildings
    - Castle
    - Fort



# Controlled Vocabularies

- Thesaurus
  - a structured wordlist used to standardize terminology to assist in indexing and retrieving information.
- It allows:
  - terms to be grouped into hierarchies and cross-referenced to other groups of terms that may be relevant to the subject.
  - the user to select a single preferred term to use
  - terms to be selected at a general or specific level
  - development by the addition, amendment and deletion of terms, relationships or hierarchies as dictated by individual needs.



# An uncontrollable world

## Users:

- Incorrectly utilizing search terms
- Unable to find what they want
- Suffer from information overload
- May as well use Google

## Creators:

- Indexing inconsistently
- Unable to convey hierarchical concepts
  - House **is a**
  - Domestic Building
  - **is a** Building
- Perpetuating localized terminology
- Unable to share data

# Gaining Control

## Users:

- Gain more effective access to a resource and *across* resources
- Reduce the number of ‘false hits’
- Find what they are looking for.
- Learn more about the resources.

## Creators:

- Produce more valuable resources
- Convey complex semantic and structural concepts
- Move towards disciplinary, national and international terminologies
- Effectively integrate both new and existing resources.

# A plethora of thesauri

- Getty Thesauri
  - Art and Architecture
  - Geographic Names
- NMR Thesauri
  - EH/RCHME Thesaurus of Monument Types
  - mda Archaeological Object Types
  - Components
  - Maritime Craft Types
  - Historic Landscape Characterization
  - Archive Types
- Other Thesauri
  - UNESCO
  - HEREIN



# Case Study: HEREIN

- Devised as a means to provide access to the national heritage policies of EU member states
- Deals with architectural and archaeological heritage, as defined in the Granada (October 1985) and Valletta (January 1992) Conventions.
- Originally conceived in English, Spanish and French



# HEREIN Thesaurus: Methodology

- Based on three basic relationships terms: equivalence, hierarchical and associative.
- Conforms to the ISO 2788: *Guidelines for establishment and development of monolingual thesauri*
- Degrees of linguistic equivalence established between the different languages, conforming to ISO 5964: *Guidelines for establishment and development of multilingual thesauri*

# HEREIN Thesaurus: Methodology

- Initially the national policy documents of Spain, Norway, Hungary, Ireland, France and the UK were ‘mined’ for terms
- These were then augmented by terms derived from national legislation and specialized documentation
- This resulted in over 1000 terms being identified in each language
- It was decided to limit the number of terms to 500 Preferred Terms to aid usability and ease of management.
- 9 thematic groups were then created to artificially subdivide the selected terms into hierarchies.
- These hierarchies were initially created in Spanish, French and English.



# HEREIN Thesaurus: Methodology

- Each term was then defined in each of the three languages and given a scope note.
- These were then translated into the other two languages.
- As a result it was then possible to define the multilingual equivalencies and the hierarchies were adjusted accordingly.

# Standard deviation...

- As the project was funded by the EU and the Council of Europe it was a stipulation that within the thesaurus there could be no source language. In this respect the thesaurus deviates from the ISO 5964 standard.
- As a result, no language is placed in a position of superiority or of prestige.
- Which in turn implies a constant exchange between languages, so as to review concepts, their translations and their relationships.

# ....and the trouble it causes!

- In a multilingual thesaurus, conforming to ISO 5964 the source language acts as a spine on which to hang all other languages.
- For example if we take French as the source language then the Spanish and English teams both simply find equivalent terms for the French term and have no need to worry about the equivalence between the Spanish and English terms.
- We then have the simple equation to determine the number of mappings required to create the thesaurus
- **t(l)** –where t = number of terms and l = number of target languages
- In this instance  $500 \times 2 = 1000$  mappings

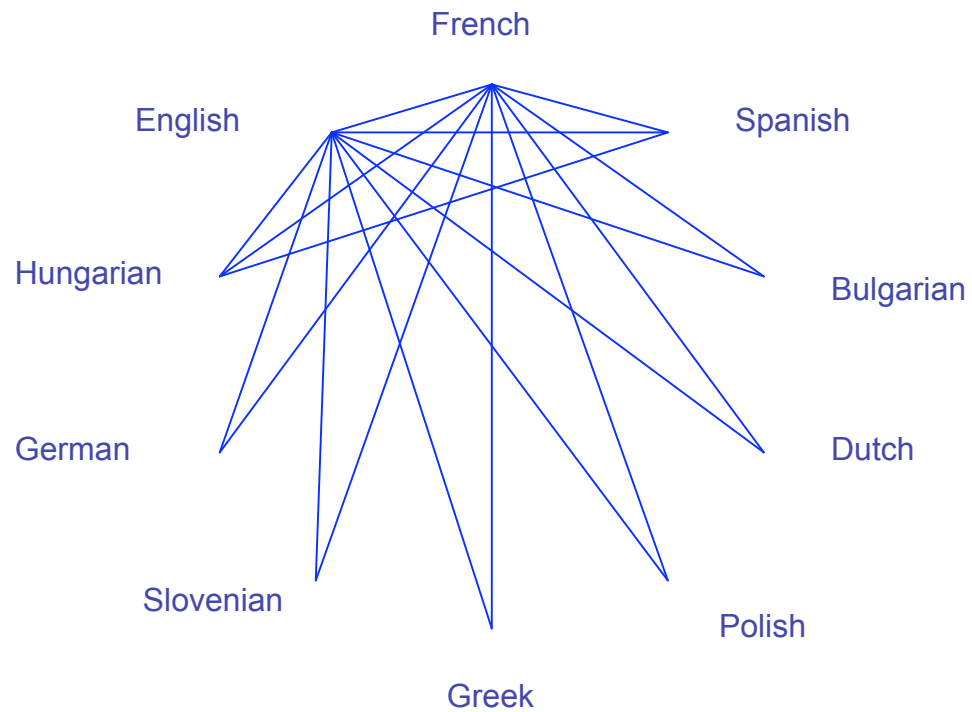
# Mapping losing it

- In the HEREIN thesaurus, each term, in each language has to be mapped to each term in every other language as the equivalence cannot be implied through the source language.
- Then the simple equation to determine the number of mappings required to create the thesaurus
- **$t(l_1 \times l_2)$**
- Where
  - **t** = number of terms
  - **$l_1$**  = number of languages
  - **$l_2$**  = number of target languages
- In this instance  $500 \times 2 \times 3 = 3000$  mappings

# Mapping gone mad

- In 2000 as part of the HEREIN 2 extension project the thesaurus was expanded to include Hungarian
  - Mappings = 6000
- By 2006 the number of languages had risen to 10 with the addition of Greek (Cypriot), Bulgarian,, Slovenian, German (Swiss), Polish, Dutch
  - Mappings = 45000
- To reduce the amount of time spent on mapping to each and every language a pragmatic decision was made in 2002 to allow new countries to only map to 2 of the original three languages.

# The working thesaurus

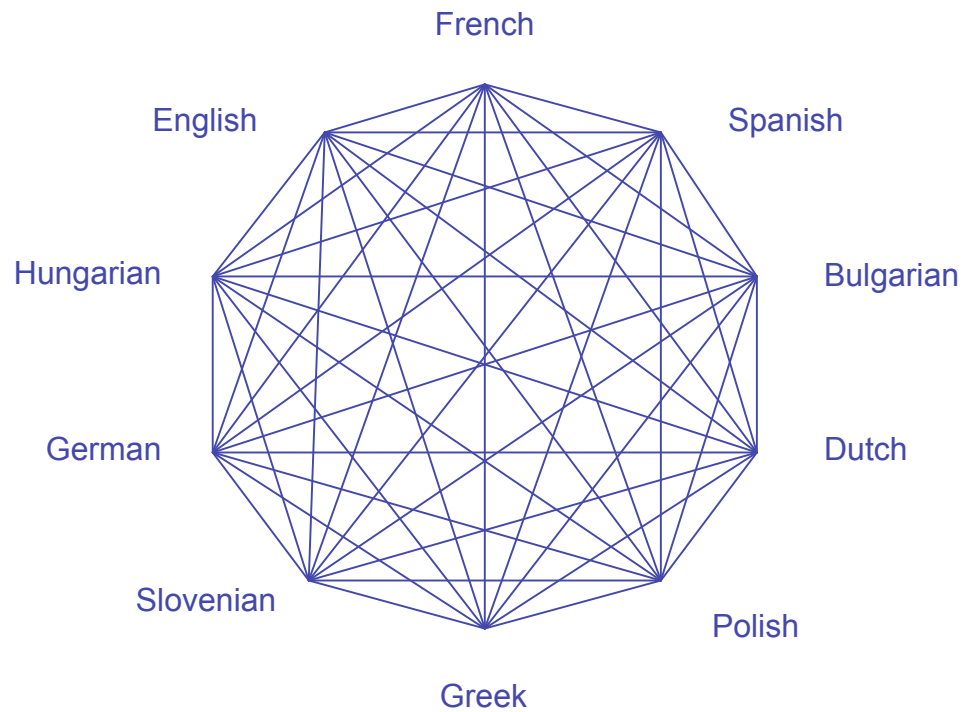


# Confused?

You: “All this talk of mapping is just nonsense! Surely the Software does all the work for you?”

Me: “What software?”

# The finished thesaurus



# The future: CIDOC CRM

- CIDOC Conceptual reference model (CRM)  
A formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information.
- The culmination of more than a decade of standards development work by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM).
- Now published as ISO 21127
- No terms – only concepts and labels
- No need for politically correct ‘equal languages’
- Enables cross-searching across systems





# Resources

- European Heritage Network

<http://www.european-heritage.net/sdx/herein/index.xsp>

- NMR Thesauri

<http://thesaurus.english-heritage.org.uk/>

- CIDOC CRM

<http://cidoc.ics.forth.gr>

- Phil Carlisle

[philip.carlisle@english-heritage.org.uk](mailto:philip.carlisle@english-heritage.org.uk)

